REMARKS

Reconsideration of this application is respectfully requested in view of the foregoing amendments and following remarks.

Response to Rejections Under 35 U.S.C. § 102

The rejection of claims 1-2, 8-12, and 18-20 under 35 U.S.C. § 102(b) as being anticipated by Sukkar (US 6292778) is respectfully traversed on the grounds that Sukkar patent does not disclose or suggest a method or system for utterance verification in which **normalization of feature vectors** using normalization parameters of the verification unit corresponding to the speech segment **is utilized** for adjusting the dynamic range of the feature vectors and generating a sequence of verification feature vectors for input to the verification-unit corresponded classifier, as recited in step (D) of claim 1. Instead, Sukkar discloses **utilization** of a **ratio** of the **likelihood** that the speech segment **contains the sound** associated with the subword hypothesis **to** the **likelihood** that the speech segment consists of a **different sound** to accordingly indicate the verification scores (col. 10, line 66 to col. 11, line 3), which is clearly different from the inventive normalization.

It is respectfully noted that the normalization recited in step (D) of claim 1 has not yet produced verification scores, and thus the claimed normalization cannot possibly correspond to the likelihood score ratio of Sukkar. In addition, the verification score of each speech segment in the invention is derived from the sequence of verification feature vectors of the speech segment, whereas

the verification scores provided by Sukkar are determined as a likelihood ratio. It is clear that the calculation for the verification scores in the invention is different from that of Sukkar.

According to the Examiner, the adjusting parameters and HMMs disclosed in col. 12, line 56 to col.13, line 18, of the Sukkar patent inherently include means and standard deviation parameters. This conclusion of inherency is respectfully traversed. In contrast to the adjusting parameters of Sukkar, which are actually model parameters in HMM and are used to estimate the likelihood of subword hypothesis, the means and standard deviation recited in claims 2 and 12 of the present application are employed to normalize the feature vectors. The adjusting parameters of Sukkar have nothing to do with such normalization.

Because claims 1, 2, 11, and 12 recites features that are different from and not anticipated by Sukkar, dependent claims 8-10 and 18-20 are also not anticipated by Sukkar, and withdrawal of the rejection under 35 USC 102(b) in view of the Sukkar patent is respectfully requested.

Response to Rejections Under 35 U.S.C. § 103

The rejection of claims 3-7 and 13-17 under 35 U.S.C. § 103(a) as being unpatentable over Sukkar (US 6292778) in view of Carey et al. (US 5526465) is also respectfully traversed on the grounds that the Carey patent, like the Sukkar patent, fails to disclose or suggest the inventive feature of using a normalization to adjust the dynamic range of feature vectors and generate a sequence of verification feature vectors for input to the verification-unit corresponded classifier.

As explained above, the method disclosed in the Sukkar patent uses the parameters to estimate the likelihood

associated with the subword hypothsis, but not to perform the normalization on the feature vectors as cited in the invention. Those skilled in the art will appreciate that the training data used for training the MLPs in the invention are pre-corrupted by noise with different power levels of SNR (for example, the speech segments corrupted by in-car noise with SNRs of 9dB, 3dB, 0dB, -3dB, and -9dB are used to train the MLPs; see page 11, lines 9-17), whereas only a certain amount of noise is given in training by Sukkar. As a result, the method of Sukkar offers poorer performance that than the MLP training provided by the invention. In order to realize the advantage of the present invention, one can use a known method (Sukkas, R.A., "Subword-based Minimum Verification Error (SB-MVE) Training for Task Independent Utterance Verification" Proc. ICASSP'98, 1998) in association with the present invention to receive noise-corrupted speech signal for implementing verification to see the difference therebetween. The result can be seen in the **supplementary document** entitled **"MLP-BASD UTTERANCE VERIFICATION FOR IN-CAR SPEECH RECOGNITION,"** which is attached hereto as **APPENDIX A**. Briefly, the invention can provide good speech recognition when the environment is changed, but the method of Sukkar cannot.

These deficiencies are not made up for by the Carey patent. In contrast to Carey, the claimed invention uses an MLP neural network as the classifier for changing the normalized feature vectors into the verification score, and not for increasing discrimination between the personal model and the world model (col. 11, lines 15-20 of Carey). In addition, Carey uses a Baum-Welch backword pass algorithm and the likelihood information in MLP training in order to increase discrimination between the personal model and the world model, whereas the claimed invention uses an error back-propagation algorithm and the information of

sequences of verification feature vectors in MLP training in order to generate the verification scores. Moreover, Carey requires two values Pp and Pw for speaker utterance training (col. 11, line 60 to col. 12, line 7), but the invention only uses the target value for speech segment training.

Accordingly, the dependent claims 3-7 and 13-17 are not suggest by the Carey and Sukkar patents, whether considered individually or in any reasonable combination, and withdrawal of the rejection under 35 USC 103(a) is requested.

CONCLUSION

In view of the foregoing remarks, reconsideration and allowance of the application are now believed to be in order, and such action is hereby solicited. If any points remain in issue that the Examiner feels may be best resolved through a personal or telephone interview, the Examiner is kindly requested to contact the undersigned attorney at the telephone number listed below.

Respectfully submitted,

BACON & THOMAS, PLLC

By: BENJAMIN E. URCIA
Registration No. 33,805

Date: May 18, 2007

BACON & THOMAS, PLLC
625 Slaters Lane, 4th Floor
Alexandria, Virginia 22314

Telephone: (703) 683-0500

MLP-BASED UTTERANCE VERIFICATION FOR IN-CAR SPEECH RECOGNITION

Shih-Chieh Chien, Tai-Hwei Hwang, and Sen-Chia Chang
Advanced Technology Center, Computer & Communications Research Laboratories, Industrial Technology Research Institute
E000 CCL/ITRI, Bldg. 51, 195-11 Sec. 4, Chung Hsing Rd. Chutung, Hsinchu, Taiwan 310
{ShihChiehChien, hthwei, chang}@itri.org.tw

Abstract

In this paper, we present the method of using Multi-Layer Perceptron (MLP) for in-car utterance verification. In this method, subsyllable-based utterance verification is conducted in our work for task independent consideration. For each subsyllable verification unit, a verification-specific MLP is employed. To avoid bias classification of using the MLP, the design of a proper set of anti-subsyllables to compete with the correct subsyllables for training is required, and a random generating procedure is introduced for this purpose. To be robust to car noise-level variation, the noise-immunity learning (Hong and Chen, 1997) is incorporated into the training of MLPs. A Mandarin digit-string verification task, simulated for the additive in-car noise, was conducted to demonstrate the effectiveness of using the MLP-based method. Experimental results show that the proposed method outperforms the HMM-based one, especially when the SNRs are low.

1 Introduction

The use of automatic speech recognition (ASR) in car environments is natural and has become one of the most promising applications of ASR. However, it is still a great challenge to achieve high accuracy for in-car speech recognition. Besides, comparing with other ASR applications, it is more difficult to correct ASR errors and the cost could be high when recognition errors occur. It is, therefore, an important issue to increase the reliability of ASR systems in car environments. Utterance verification (UV) that is able to detect and reject recognition errors and noise tokens is a promising technology for this purpose.

Many methods for UV have been proposed in the last decade (Sukkar and Lee, 1996; Pao *et al.*, 1998; Sukkar, 1998), but seldom of them are developed for noisy conditions. To apply UV for car applications, the most concerned issue is the way to handle the mismatch between the training and the test sets to avoid performance degradation. This mismatch is especially serious for the wide range and time varying of in-car noise since it is difficult to design a well solution to handle all practical conditions. As an alternative, Hong and Chen (1997) proposed an RNN-based immunity learning procedure to train the classification models that can be adapted to match/mismatch noisy conditions for front-end pre-classification. In this learning, the corrupted training material, corrupted by background noise with gradually decreasing the SNR levels, is used to train the model for obtaining the noise immunity. They also suggested using the neural networks instead of the HMMs for the estimation of noise statistics. Base on their study, the same technique is applied to UV for in-car environment, more specifically, for reducing the effect of additive in-car noise. And the multi-layer perceptron (MLP) is utilized to this task.

The MLP, which has been widely investigated in many fields, can be considered as another way of computing a verification measure. In the study of Modi and Rahim (1997), it was applied to integrate multiple confidence measures for minimizing the verification errors. In this paper, the MLP is used to estimate the confidence score of a given subsyllable segment by feeding the same features used for recognition directly and, more important, to make UV robust to car noise-level variation. Taking this approach, it would be beneficial to incorporate with the speech recognition using of the subsyllable-based models for developing in-car speech applications. To reach this goal, the speech

segments of subsyllable and anti-subsyllable corrupted by car-noise are used for the MLP training. The training data design concerning the effectiveness of using this approach is the main subject to this work. The selection of training patterns to avoid bias classifications and the arrangement of training data for immunity learning are disclosed in this paper.

The remainder of the paper is organized as follows. In Section 2, the architecture of the MLP-based UV and the issue concerning about the selection of training patterns are discussed. The immunity learning of MLP is presented in Section 3. In Section 4, an in-car Mandarin digit-string verification task was performed to evaluate the effectiveness of using the proposed method. A brief conclusion is given in the last section.

2 MLP-based utterance verification

In the study of Sukkar and Lee (1996), utterance verification is treated as the problem of statistical hypothesis testing. Where the null hypothesis is tested against the alternative hypothesis and the resulted likelihood ratio is used to indicate the confidence score of the verification target. Utterance verification also could be regarded as a two-class classification problem, i.e., to classify whether the verification target is a member of correct-class or incorrect-one. In this paper, we applied MLPs to solve this two-class classification problem.

2.1 Subsyllable-based MLP verifier

We depicted the subsyllable-based verification by using MLPs in Figure 1. In this verification procedure, the test utterance was recognized and then segmented into several subsyllable segments according to the lexical structure of the recognized hypothesis. For each subsyllable segment, a subsyllable-level verification score is generated by a corresponding MLP. The resulted utterance-level confidence score is then obtained through combining the verification scores of these subsyllable segments.

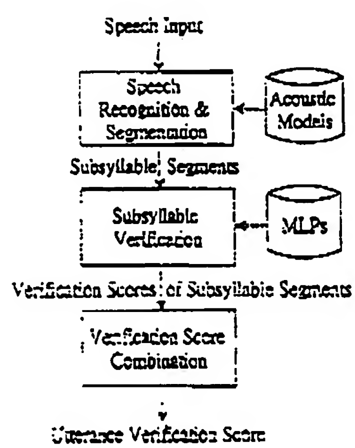


Figure 1: Utterance verification block diagram.

The MLP, used for the subsyllable-based verification, is a three-layer architecture with single output node for generating the classification score. The input layer accesses the acoustic feature vectors of speech frames and feeds them forward to acquire the classification score. This process is performed on each frame of subsyllable segment. The subsyllable-level verification score is obtained when the feed-forward processes on subsyllable segment are finished.

The error back-propagation training algorithm is used in the MLP training. To achieve the goal of two-class discrimination, the subsyllable segments corresponding to correct-class (or correct subsyllable) and incorrect-one (or anti-subsyllable) are used as the training materials. In this training

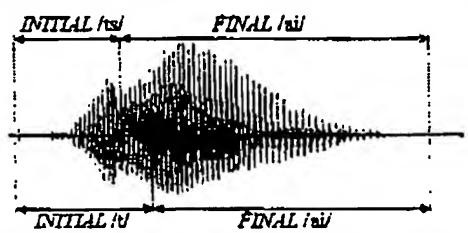
procedure, the criterion of minimum mean-square error is used to minimize the actual and the desired output scores and is referenced to back-adjust the free parameters of the MLP. The desired output of 1 and 0 is set to indicate the verification score of the correct subsyllable and the anti-subsyllable respectively.

2.2 Training data design

To train the subsyllable-based verification model by using MLP, the correct and incorrect (or anti) subsyllables are included in the training material for discrimination. In general, the incorrect part would have to represent all of the correct part's complements. However, the acoustic confusion always can be found between the target subsyllable and its complements, the biased classification is inevitable if all of the complements are used to compete with the target subsyllable in the MLP training. This problem is not only due to the acoustic confusion between subsyllables but also derived from the unbalanced amount of competitors. Therefore, the quantity control and a general representation of anti-subsyllables are the key-issue to this problem, and the "random lexicon", introduced by Sukkar and Lee (1996), is used. The random lexicon used in their work simulates the incorrect recognition for verification result observation. In our study, it is utilized to generate a similar amount of anti-subsyllable patterns with respect to the correct subsyllable patterns for the MLP training. The collection of the training patterns is listed in the following steps.

Step-1 Transform the training text into syllable-denoted sequence and denote it the correct syllable sequence.
Step-2 Perform subsyllable-level segmentation with the correct syllable sequence on training speech and collect the subsyllable segments for the correct-part.
Step-3 Duplicate the correct syllable sequence and name the duplication the operating sequence.
Step-4 Randomly cut the syllable in the operating sequence and paste it to a new syllable sequence. During this operation, the syllable at the position of the new sequence must be kept different from the one with the same position of the correct sequence. This new syllable sequence is denoted the random syllable sequence.
Step-5 Perform subsyllable-level segmentation with the random syllable sequence on the same training speech and collect the subsyllable segments for the anti-part except the segment with the same subsyllable denotation as the correct one.

Correct syllable "ㄗㄞ" (represented by *INITIAL* /ts/ and *FINAL* /ai/)



Random syllable "ㄊㄞ" (represented by *INITIAL* /t/ and *FINAL* /ai/)

Figure 2: The correct and random syllables.

We depict an example in Figure 2 to clearly describe the collection of anti-subsyllables. In this figure, a random syllable "ㄊㄞ" (represented by *INITIAL* /t/ and *FINAL* /ai/) and the correct syllable "ㄗㄞ" (represented by *INITIAL* /ts/ and *FINAL* /ai/) is located at the same position of the random and correct syllable sequence, respectively. The correct (/ts/ and /ai/) and random (/t/ and /ai/) subsyllable segments are obtained after performing speech segmentations with the correct and random syllable structures on this utterance, respectively. The subsyllable segment of /ts/ and /ai/ is collected separately for the correct data of subsyllable /ts/ and /ai/. For anti-subsyllable collection, the segment of /t/ is regarded as the

anti-data of subsyllable /t/. However, the other one, /ai/, is ignored since it is identical to the correct one in denotation and, partially, in acoustics.

After performing segmentation with the random syllable sequence, the anti-subsyllable segments are obtained and represented by the most likely segments in the training utterance, but they have different characteristics from the correct subsyllables. These anti-subsyllable segments can be regarded as the incorrect recognition results and should be rejected by system. And they are used to compete with the correct subsyllable segments in training. Furthermore, the random sequence generation (in Step-4) can be also regarded as a rearrangement process on the correct syllable sequence. The total syllables in the random syllable sequence are the same as in correct syllable sequence. In consequence, a similar amount of the correct and incorrect patterns can be collected from Step-2 and Step-5. Although some incorrect patterns are dropped in Step-5, the abandoned quantity can be ignored with respect to the final quantity of incorrect data in our implementations. Therefore, it guarantees us against the biased classifications using of the balanced data for the MLP training.

3 Robust MLP training

To apply speech technologies for in-car applications, the capability of handling car noise variation is the most concern issue. The immunity learning proposed by Hong and Chen (1997) provides an alternative to tolerate the wide range of noise variations. To obtain noise immunity, the speech signals corrupted by background noise with several noisy conditions are used in their training. And an RNN-based learning scheme is utilized for this method. The RNN, which mounts feedback tentacles to transfer previous information for more detailed estimations, is a suitable mechanism to model the dynamic change of speech patterns. Therefore, in their implementation, the speech signals corrupted by different noisy conditions are sequentially used to train the RNNs.

Based on the same learning technique, speech signals infected with noise are also used for the MLP training. Different from the RNN-based training, the "clean" training data of subsyllable and anti-subsyllable is duplicated into several copies for different corruptive conditions and all of them are used to train the MLP at a time. That is, we expect the MLP training not only to learn the discrimination capability of two-class but also to tolerate the change of environmental conditions.

4 Experiments

Effectiveness of the MLP-based UV is examined by simulations on a Mandarin digit-string recognition task. To simulate the in-car environment, the data area of NTT-AT database (NTT-AT, 1996) was used. And we chose the in-car noise of CIVIC for the target scenario in this experiment.

4.1 Recognition models

The ASR used for this task is an HMM-based speech recognizer. We employed 20 subsyllable models, including 10 three-state *INITIAL* models and 10 five-state *FINAL* models, for recognizing the 10 Mandarin digits. These HMMs were trained by using the digit-string database of MAT (Wang, 1997), which is designated as the "clean" speech data.

4.2 The Databases for verification model training

The same digit-string database for training the recognition models was also used to train the verification models. The training set includes 4688 utterances and 2504 speakers. The length of each digit-string is ranging from 4 to 7 digits. The development set, including 1159 utterances spoken by 1159 speakers, was used for selecting the verification models with optimal performances in the iterative training results. The contents of this part are all 7 digits.

For robust training, two in-car noisy conditions of CIVIC, the in-car noise of high-speed driving (CIVIC-1) and low-speed driving (CIVIC-2), are used. And these two noisy conditions coordinated with the SNRs (in the power levels of speech and noise signals) of 9dB, 3dB, -3dB, and –9dB were artificially add to the speech signals of training set for noise-immunity learning. The same noisy conditions were

also used to corrupt the development set but using the SNRs of 6 dB, 3 dB, 0 dB, -3 dB, and -6 dB for the optimal model selection.

4.3 Verification models

The feature vector used for recognition and verification consists of 26 coefficients, including 12 mel-cepstral coefficients, 12 delta mel-cepstral coefficients, 1 delta energy, and 1 delta-delta energy. In the constructions of MLPs, 78 neurons are designed in the input layer for accessing the acoustic feature vectors of three speech frames. Besides the single output neuron in the output layer for generating the classification score, 30 neurons are used in hidden layer to arbitrate between the input and the output of neural network. To calculate the verification score of a speech segment, three speech frames are fed into the MLP to obtain a classification score for every frame-slot except the first and the last frame-slot, that is, there are (T-2) classification scores for a speech segment with T frames. And the verification score of this speech segment is the mean value of the (T-2) classification scores.

We followed the descriptions in section 2 and section 3 to design the noise-infected data and train the MLPs, where 150 training iterations were used in the training. The optimal parameter-set was select from the 150 training results, and the equal-error-rate (EER) of false rejection and false alarm is used as the selection criterion. Finally, we obtained 20 MLPs corresponding to the 20 recognition HMMs for the MLP-based verification.

The models for the HMM-based utterance verification were also trained in the same experimental conditions for comparative study. The MVE training (Sukkar, 1998) was performed to train the subsyllable-based verification models, and we had 20 subsyllable HMMs and 20 anti-subsyllable HMMs for the HMM-based verification.

4.4 Experimental evaluation and discussion

A pre-defined phone book, including 68 telephone numbers, is the target for recognition and verification in our evaluation. In order to exam the verification performance, we describe the in-vocabulary (IV) and the out-of-vocabulary (OOV) databases used in this evaluation as follows.

- DB1: The IV database. This database was collected through telephone networks. There are 1948 utterances spoken by 661 speakers. The digit-length is 6 or 7 in the pre-defined phone book.
- DB2: The first set of OOV database, which was recorded by handset microphone. This database is also a digit-string database but the contents are not included in the phone book. The length of each digit-string is ranging from 2 to 6 digits. This database includes 99 speakers and 504 utterances.
- DB3: The second set of OOV database, which was collected from our auto-attendant system (Jou *et al.*, 2000) through telephone networks. The Chinese person-names are the contents of this database. This database contains 3479 utterances.
- DB4: The third set of OOV database, which contains the noise data and the spontaneous speeches, such as "um", "ah", and some incomplete sentences collected from our auto-attendant system. We collected 1219 utterances for this set.

These 4 sets of databases were also corrupted by adding the CIVIC-car noise but with the SNRs of 6 dB, 0 dB, and –6 dB separately to observe the robustness of verification models.

We performed utterance verification after digit-string recognition to reject the recognition error. The verification results were summarized and plotted in Figure 3. Again, the EER is used as the evaluation criterion for performance observations. It can be observed from this Figure that the verification difficulty is increasing as the SNRs are decreased. And the verification is more difficult for the noisy condition in high-speed driving (CIVIC-1) than the low-speed one (CIVIC-2). It also can be understood from this Figure that the non-digit-string databases, DB3 (person-names) and DB4 (noise and spontaneous speech), are easier to reject than DB2 (digit-string).

Examining the performances of these two verification methods, the EERs of the HMM-based verification are almost in linear increasing as the SNRs are decreased from 6 dB to –6dB. The EERs of

the MLP-based method are quite equal when operated at 6 dB and 0 dB. But there is a little bit high for the –6 dB condition. To be more clearly, we further summarized these results in Table 1. It can be noted from this table that the performance of the MLP-based method and the HMM-based one is very similar for the 6dB case. On the other cases, the MLP-based verification presents better results than the HMM-based one, especially for the –6dB condition.

For the rejections of substitution errors, we listed the evaluation results of different test conditions in Table 2 for the baseline system and the system incorporated with UV. The substitution errors are decreased largely when 3% of false rejection is set. This evaluation also shows the better result on the reduction of substitution errors by using the MLP-based verification. From these results, it can be concluded that the MLP-based verification is more insensitive to noise-level variation than the HMM based method. This conclusion also coincides with the result of Hong and Chen (1997).

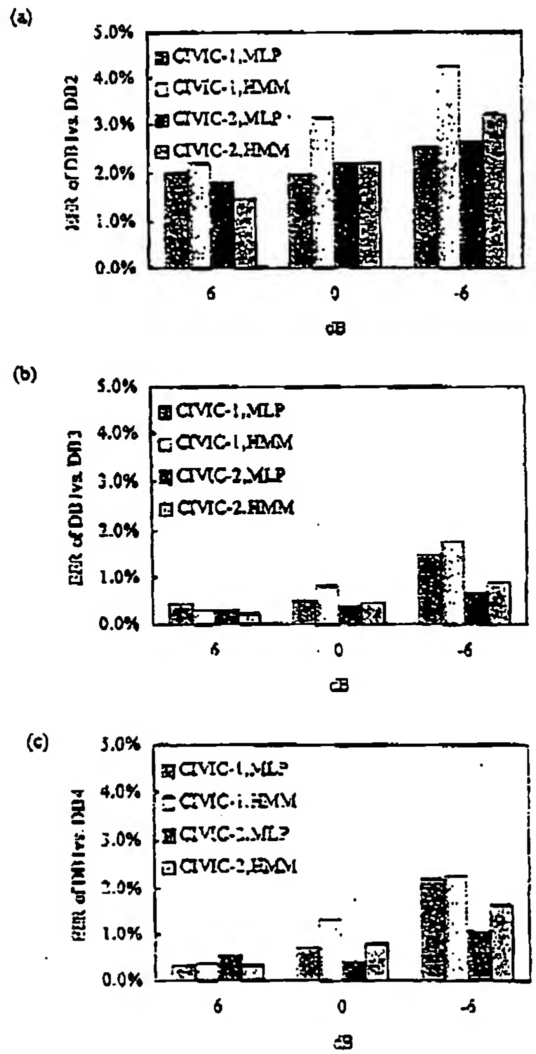


Figure 3: Utterance verification results for different nosy conditions. Where the IV database, DB1, was tested against the OOV databases of (a) DB2, (b) DB3, and (c) DB4 .

Table 1: The averaged verification performances of the MLP-based method and the HMM-based one for the SNRs of 6 dB, 0 dB, and –6 dB.

SNR (dB)	Averaged EER of	
	MLP	HMM
6	0.89%	0.81%
0	1.02%	1.45%
-6	1.78%	2.34%

Table 2: The substitution errors for the system with and without utterance verifications under different noisy conditions.

Noise Type	SNR (dB)	Baseline system		With verification		
		False Rejection	Substitution Error	False Rejection	Substitution Error	
					HMM	MLP
CIVIC-1	6	0%	1.44%	3%	0.31%	0.21%
	0	0%	2.10%	3%	0.31%	0.46%
	-6	0%	8.88%	3%	3.44%	2.82%
CIVIC-2	6	0%	1.28%	3%	0.36%	0.26%
	0	0%	1.54%	3%	0.41%	0.31%
	-6	0%	3.34%	3%	0.87%	0.72%

5 Conclusion

This paper presented the MLP-based utterance verification for in-car speech recognition. In this method, the subsyllable-based verification models were used. To avoid bias result of using the MLP, a balanced amount of subsyllable and anti-subsyllable patterns and a general representation of anti-subsyllables for discriminative training are required. A procedure for the collection of training patterns was introduced for this requirement. To be robust to car noise-level variation, the noise-immunity learning was applied and implemented by using the car-noise-infected speech as the training materials of MLPs.

From our experimental results, we found that this MLP-based method is more insensitive to noise-level variations than the HMM-based one. It seems to be a workable method for in-car utterance verification. However, the study on real in-car environment is still needed since only additive noise is considered in this study. Moreover, encouraged by these results, scaling up this approach to large vocabulary recognition will be also considered in our future work.

Acknowledgements

This paper is a partial result of project B32BXT5100 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan.

References

Hong, W.-T. and S.-H. Chen. 1997. A Robust RNN-based Pre-classification for Noisy Mandarin Speech Recognition. *Proc. Eurospeech'97.*

Jou, S.-C., S.-C. Chien, W.-C. Shieh, J.-H. Chen, and S.-C. Chang. 2000. CCL eAttendant - An On-line Auto-attendant System. *International Symposium on Chinese Spoken Language Processing (ISCSLP) 2000.*

Modi, P. and M. Rahim. 1997. Discriminative Utterance Verification Using Multiple Confidence Measures. *Proc. Eurospeech'97.*

NTT-AT. 1996. *Ambient Noise Database for Telephonometry 1996.* http://www.ntt-at.com/index.html, NTT Advanced Technology Corp..

Pao, C., P. Schmid, and J. Glass. 1998. Confidence Scoring for Speech Understanding Systems. *Proc. ICSLP'98.*

Sukkar, R.A. 1998. Subword-based Minimum Verification Error (SB-MVE) Training for Task Independent Utterance Verification. *Proc. ICASSP'98.*

Sukkar, R.A. and C.-H. Lee. 1996. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition. *IEEE Trans. On Speech and Audio Proc.*, 4(6): 420-429.

Wang, H.-C. 1997. MAT – A Project to Collect Mandarin Speech Data Through Telephone Networks in Taiwan. *Computational Linguistics and Chinese Language Processing*, 2(1):. 73-90.

Proceedings of

Oriental COCOSDA 2003

International Coordinating Committee on Speech Databases and Speech I/O System Assessment

1-3 October, 2003
Sentosa, Singapore

Edited by
Min Zhang, Haizhou Li and Kim Teng Lua

COLIPS PUBLICATIONS